CDA LEVEL Ⅲ 考试大纲

CERTIFIED DATA ANALYST LEVEL Ⅲ EXAMINATION OUTLINE

CDA 考试大纲是 CDA 命题组基于 CDA 数据分析师等级认证标准而设定的一套科学、详细、系统的考试纲要。考纲规定并明确了 CDA 数据分析师认证考试的具体范围、内容和知识点，考生可按照 CDA 考试大纲进行相关知识的复习。

数据治理（占比 35%）

a. 数据治理概述、大数据类型（占比 5%）
b. 元数据管理、数据标准（占比 7%）
c. 大数据隐私、安全 （占比 5%）
d. 数据质量管理（占比 8%）
e. 数据生命周期管理（占比 5%）
f. 数据服务（占比 5%）

大数据架构设计（占比 35%）

a. 架构设计方法论概述（占比 5%）
b. hadoop 生态体系（占比 10%）
c. 大数据分层架构设计思想（占比 5%）
d. 大数据之常见场景下的架构设计（占比 15%）

机器学习（占比 30%）

a. 线性模型（含 Logistic 回归、偏最小二乘法 PLS、Lasso 回归等）（占比 3%）
b. 决策树（占比 2%）
c. 集成学习提升方法（占比 1%）
d. 随机森林（占比 5%）
e. 最近邻方法 KNN（占比 1%）
f. 朴素贝叶斯（占比 1%）
g. 支持向量机 SVM（占比 3%）
h. 感知机和神经网络（占比 3%）
i. 极大似然估计和 EM 算法（占比 1%）
j. K-means 聚类（占比 1%）
k. 关联规则（Apriori 算法）（占比 1%）
l. PCA 降维（占比 1%）
m. 大规模机器学习（占比 3%）
n. 深度学习（占比 3%）
o. 模型的评估与选择（占比 1%）

CDA LEVEL Ⅲ数据科学家考试大纲解析

CERTIFIED DATA ANALYST LEVEL Ⅲ EXAMINATION NOTE

根据 CDA 数据分析师认证考试大纲，经管之家 CDA 数据分析研究院给出了详细解析，以“领会”，“熟知”，“应用”三个不同的级别将每一个知识点进行分解，建议考生应该按照不同的知识掌握程度有目的性的进行复习。

1. 领会：要求应考者能够记忆规定的有关知识点的主要内容，并能够了解规定的有关知识点的内涵与外延，了解其内容要点和它们之间的区别与联系，并能根据考核的不同要求，做出正确的解释、说明和阐述。
2. 熟知：要求应考者必须熟悉的理论知识，并能够正确理解和记忆相关的理论方法，根据考核的不同要求，做出逻辑严密的解释、说明和阐述。
3．应用：要求应考者必须掌握知识点的主要内容，并能够结合工具进行商业应用，根据考核的具体要求，做出问题的具体实施流程和策略。

PART 1

数据治理

- 数据治理概述、大数据类型

1．领会：数据治理的概念、框架、目标；数据治理能力成熟度模型（DMM）及分级实施；

2．熟知：了解大数据的不同类型及其处理方法，如 web 和社交网络或媒体数据、机器产生的数据、人工生成的数据、高频交易数据、生物统计数据等；结构数据与非结构数据；内部数据与外部数据等。大数据与主数据的协同。

- 元数据管理、数据标准

1．领会：元数据的概念、存储、管理；数据标准的概念

2．熟知：元数据的编码体系：语法编码和受控词汇表；都柏林核心元数据元素集；基于业务元数据、技术元数据、操作元数据间的关联关系，构建元数据模型；了解元数据管理自动化；熟悉业务词库的建立和数据标准的实施

3．应用：建立企业级元数据管理系统；建立企业级元数据技术规范和实施指引；制定合适的元数据管理流程；

- 大数据隐私、安全

1．领会：个人可识别信息 PII，识别敏感的大数据；熟悉国家隐私相关法律要求；

2．熟知：对元数据库中的敏感大数据进行标记；隐私影响评估 PIA；监控特权用户对敏感大数据的访问；管理个人数据跨国界流动的情况

3．应用：隐私保护措施的实施，包括从设计入手保护隐私；对数据采集、保留、处理的合理限制；获得用户的明示同意；对数据进行反识别；要求下游用户将数据以反识别的形式保存等。物理安全、系统安全、存储数据安全的实施，其中系统涉及技术包括 hadoop, mapreduce, nosql 等；数据加密算法，如对称加密 DES，IDEA；非对称加密 RSA 算法、ECC 算法；散列算法；数字签名；数字证书等。

- 数据质量管理

1．领会：数据剖析在大数据应用中的作用；数据处理理论；数据质量评测方法；数据质量管理平台基本功能及构架；

2．熟知：数据剖析方法与原理；数据剖析的基本流程；剖析中的业务规则的应用；数据质量诊断的原理和方法；数据诊断流程；数据质量诊断报告及解读；数据治理中的业务规则定义及使用；企业业务规则库建立及使用；企业数据质量管理平台使用角色及职责权限设计；数据质量管理平台在不同的商业应用场景的定位及部署；

常用质量管理工具，如帕累托图、鱼骨图、休哈特图、智能设备校准；

3．应用：数据质量管理工具与 ETL、数据分析工具的区别与交互作用；数据库查重；

- 数据生命周期管理

1．领会：大数据生命周期管理概念；对数据热度的理解，如热数据、温数据、冷数据；数据整合与主数据管理；

2．熟知：对大数据生命各周期进行管理，如定义大数据范围、大数据采集、大数据存储、大数据整合、大数据呈现与使用、大数据分析与应用、大数据归档与销毁；

3．应用：数据实时采集、抽取技术；非结构化数据采集技术；数据存储及清理技术；数据可视化技术；

- 数据服务

1．领会：数据地图、指标库、客户信息统一视图概念；

2．熟知：大数据在个性化客户服务、市场预测、用户体验、风险管控、精准营销、舆情分析、经营管理等方面的应用

3．应用：大数据整合与集成业务数据，联通数据孤岛

PART 2

大数据架构设计

- 架构设计方法论概述

1．领会：管理架构视图、业务架构视图、分层架构设计的思想和技术架构视图的概念及涉及范围

2．熟知：架构文档管理、架构风险管理模型、业务需求分析的常用工具、总体架构设计的工具和方法、分层架构设计的过程和内容。

3．应用：能运用架构设计的方法体系进行企业信息化架构设计的实现。

- hadoop 生态体系

1．领会：hadoop 分布式计算框架的思想、Google 分布式计算的三驾马车

2．熟知：深入理解 Hadoop 体系架构，包括 HDFS、MapReduce 和 Hbase 的原理、机制和优化配置过程

3．应用：能搭建基于 hadoop 体系的大数据平台，并进行参数调优

- 大数据分层架构设计

1．领会：大数据分层架构的思想，大数据的五大分层逻辑结构、基础框架设计思想

2．熟知：数据分析层的内容及相关组件、编程模型层的内容及相关组件、数据存储层的内容及相关组件、文件存储层的内容及相关组件、数据集成层的内容及相关组件、服务总线设计

3．应用：可以根据企业的业务特征，设计出基于五大分层思想的逻辑架构设计。

- 大数据之常见场景下的架构设计

1．领会：各种大数据分析场景下，模型的选择；海量存储、离线计算、在线计算、流式计算等四种常见的大数据分析场景的区别与联系

2．熟知：HDFS、GFS、fastDFS 等常用海量存储工具，mapreduce、hive、dremel、drill、Impala 等离线计算工具，BigTable、Hbase、SimpleDB、DynamoDB、Redis、MongDB 等在线计算工具，S4、Storm、Sanza 等流式计算工具，Zookeeper、Pregel、Mahout、Spark、Caffe、Kafka 等常用大数据工具。

3．应用：能设计出大数据之离线计算、在线计算、流式计算等常见的业务场景下的架构。

PART 3

机器学习

- 线性模型（含 Logistic 回归、偏最小二乘法 PLS、Lasso 回归等）

1．领会：线性模型建模的思想，Logistic 回归建模的参数估计的极大似然方法，偏最小二乘法 PLS 的原理、Lasso 回归建模的原理。

2．熟知：Logistic 回归建模的 OR 值和回归系数之间的关系，偏最小二乘法 PLS 的应用场景，Lasso 回归优化模型的方法

3．运用：正确进行 Logistic 的参数估计和预测值、概率值之间的换算，运用 Logistic 模型进行预测和模型评估，会运用偏最小二乘法进行满意度的计算，运用 Lasso 回归进行变量筛选(Variable Selection)和复杂度调整(Regularization)。

- 决策树

1．领会：决策树数据分析和预测的思想，不同决策树算法 CART、C4.5 的算法思想，领会 CART 算法，领会决策树剪枝的原理。

2．熟知：决策树（含回归树）不对数据类型的建模分析过程和预测方法，Gini 不纯度、熵、错误率的计算方法，决策树剪枝的方法。

3．运用：能使用 ID3、CART 和 C4.5 算法对不同的数据集进行决策树建模，并给出准确

率的估计，能通过相应的算法给出决策树的正确结果，能借助 R 和 Python 软件进行剪枝优化。

- 集成学习提升方法

1．领会:机器集成学习的原理及 Bootstrap 的原理,Bagging 降低方差的原理,AdaBoost 增大错误样本权重，减小正确样本权重的原则。GDBT 结合梯度迭代和回归树的方法。

2．熟知： 集成学习投票法的过程和输入输出项的含义，bootstrap 方法的 R 语言和 Python 的实现，集成算法中的方差和偏差。

3．应用： 会使用集成学习方法进行图像识别和特征选择，会使用集成方法对给定的数据集进行有监督的学习和预测，会使用 ROC 曲线判别识别的精度，完成对模型预测准确度的评价。

- 随机森林

1．领会：随机森林的训练过程和预测过程，CART 和 Gini 值在随机森林中的作用。

2．熟知:随机森林算法和集成学习算法的区别和联系，随机森林进行特征选择的方法，随机森林填补缺失值的方法。

3．应用：借助于 R 和 Python 软件，对给定数据集完成特征选择（feature selection）和预测，并能对其中的关键参数如决策树的个数和变量的个数进行参数调优，得到最佳的随机森林模型。能够结合概率和图形分析随机森林的计算结果。

- 最近邻方法 KNN

1．领会：最近邻方法在分类预测和推荐系统中的算法原理，领会计算距离的各种方法，领会最近邻方法和协同过滤等方法的比较优势。

2．熟知：最近邻方法的算法过程。

3．应用： 借助 R 语言或者 Python，能灵活运用最近邻方法到各种有监督的分类模型，包括对推荐系统的构建，并能准备评价其预测的准确率和召回率等，并能和其它有监督预测方法进行比较。

- 朴素贝叶斯

1．领会：朴素贝叶斯的算法原理和流程，离散型和连续型贝叶斯公式的应用。

2．熟知：朴素贝叶斯方法的算法过程。

3．应用：借助 R 语言或者 Python，能灵活运用朴素贝叶斯方法到各种有监督的分类模型，能够利用朴素贝叶斯进行文本挖掘、检测不真实账号、预测垃圾邮件，对朴素贝叶斯分类器进行评价。

- 支持向量机（SVM）

1．领会：支持向量机算法实现的三重境界，领会线性可分和线性不可分的概念，领会核函数的原理和功能，领会超平面的概念，领会间隔及其函数的含义。

2．熟知：支持向量机在 R 和 Python 中的建模和预测的方法和过程，熟知支持向量机的 SMO 算法。

3．应用：借助 R 语言或者 Python，能结合给定的数据集，在完成对其的数据清洗后，能使用支持向量机建模并进行 4 个重要的参数调优，能使用支持向量机进行各种数据的分类和文本分类。

- 感知机和神经网络

1. 领会：感知器的网络结构和神经网络的神经元的原理，领会感知器的学习规则和网格训练；领会两者的权重和阈值概念，领会神经网络的算法原理。
2. 熟知：感知器的计算过程和神经网络的 sigmoid 的原理和函数，熟知神经网络迭代次数的控制，熟知神经网络过度学习的现象。
3. 应用：借助 R 语言或者 Python，能结合给定的数据集，在完成对其的数据清洗后，正确的借助软件完成对数据的输入和输出，并能控制其中的核函数，得到比较准确的预测结果，并完成对结果的评价。会绘出神经网络的精度折线图，并能控制好迭代和学习的精度以不至于过度学习。

- 极大似然估计和 EM 算法
  1. 领会：极大似然估计和 EM 算法的原理，领会似然函数取极值的思想，领会 EM 算法中期望和极大似然估计之间的联系，领会高斯混合分布及其模型。
  2. 熟知：用极大似然估计进行参数估计的各种情形，包括对离散数据和连续数据分布的参数估计；熟知 EM 算法隐藏变量的概念，熟知 E 步和 M 步的算法具体公式和过程 。
  3. 应用： 熟练使用极大似然估计的方法进行各种分布的参数估计；能准确找出 EM 算法的隐藏变量和概率方法，能完成各种情形下的 EM 算法的参数估计。
- K-means 聚类
  1. 领会：K-means 聚类的算法原理，领会无监督学习的思想。
  2. 熟知：K-means 算法的步骤，R 语言或 Python 实现的 K-means 聚类的结果解释。
  3. 应用：能使用 R 语言或 Python 进行各种数据的 K-means 聚类，能科学确定其中的 K 的取值，能使用工具进行中位数的聚类。关联规则（Apriori 算法）
- 关联规则（Apriori 算法）
  1. 领会：关联规则（Apriori 算法）的算法原理，领会使用候选项集找频繁项集的思想。
  2. 熟知：频繁项集、支持度、置信度和提升度的概念，熟知生成频繁项集的 Apriori 算法的基本过程。
  3. 应用：能使用 R 语言进行购物篮分析、泰坦尼克号等经典数据的分析，并能确定最佳的关联规则，能对冗余的关联规则进行准确删除，能合理解释关联规则的结果，能初步使用 Python 写出 Apriori 的算法。 PCA 降维
- PCA 降维
  1. 领会：主成分分析（PCA）的思想和计算过程的协方差矩阵的作用，领会 PCA 的应用范围的数据压缩和数据可视化。
  2. 熟知：主成分分析（PCA）的特征根和特征向量的计算，PCA 的贡献度的计算，PCA 的各种软件的实现过程和结果解释。
  3. 应用：能使用各种工具对图像进行 PCA 降维，并完成图像的重构；能正确确定最佳的主成分个数并提取主成分；能使用 PCA 方法完成综合评价；能用计算机语言一步步写出 PCA 的计算过程。
- 大规模机器学习
  1. 领会：支持向量机的大规模机器学习算法，领会神经网络和深度学习在大规模机器

学习中的应用，领会 Hadoop 生态系统和 Spark 进行大规模计算的架构方式和计算原理，领会并行计算的原理。

2. 熟知：熟练运用 Python 的 pandas（I/O）工具，熟知批梯度下降和随机梯度下降的算法，熟悉正则化特征选择的方法，熟练掌握 XGBoost 算法，熟练掌握“out-Of-core”的 CART 算法，熟悉 Spark 数据压缩、数据清洗、交叉验证的方法，熟知实时计算和并行计算的过程。
3. 应用：能构建“out-of-core”学习系统，能使用 SDG 进行数据流的特征管理，能进行非线性的支持向量机的子抽样算法和梯度下降算法，能使用 Python 结合 Vowpal Wabbit（超参数调优）进行支持向量机的快速大规模学习，能够使用 TensorFlow、SkFlow 和 Keras 进行大规模机器学习，能够进行 GPU 和 Theano 计算，掌握大规模机器学习的 PCA 算法和 K-means 算法。

- 深度学习

1. 领会：基于梯度的学习、隐藏单元、反向传播、范数惩罚、多任务学习、稀疏表示、集成深度学习的原理，领会循环神经网络和递归网络，深度学习中的结构化概率模型和生成模型。
2. 熟知：（随机）梯度下降算法、动量算法、优化策略和元算法。熟悉卷积神经网络的卷积算法，熟知参数优化算法，熟知线性因子模型。
3. 应用：熟练掌握概率分布和主成分分析 PCA 在深度学习中的应用，熟练掌握极大似然估计、贝叶斯估计、（无）监督学习的各种算法在深度学习中的应用，掌握 Bernoulli 输出分布的 Sigmoid（含 logistic）单元，掌握 BP 的计算方法，熟练掌握深度学习的正则化技术，掌握具有自适应学习速率的算法，能用卷积神经网络、循环网络和递归网络进行深度学习构建。

- 模型的评估与选择

1. 领会：交叉验证（Cross-Validation）、错误率、精度、召回率、ROC、AUC、假设检验的思想。
2. 熟知：交叉验证（Cross-Validation）在 R 语言和 Python 中的实现方法，熟知各种情形下错误率、精度、召回率、ROC、AUC 的计算。
3. 应用：能够对交叉验证的结果给予解释，对预测结果会计算其错误率、精度、召回率等并完成其评价，熟悉 ROC 曲线的绘制，能够对各种结果进行假设检验并得到结论。能够综合上述的分析结果完成对模型的评估并挑选出最佳的预测模型。

参考书目

- 元数据-用数据的数据管理你的世界，[美] 杰弗里•波梅兰茨 （Jeffrey Pomerantz） 著；李梁 译， 中信出版集团
- 大数据治理，[美] 桑尼尔•索雷斯（SUNIL SOARES）著；匡斌 译， 清华大学出版社
- 架构之美 Till Adam 著；王海鹏 / 蔡黄辉 / 徐锋译；机械工业出版社，2009
- 《Hadoop 权威指南（第 3 版 修订版）》，Tom Wbite 著；华东师范大学数据科学与工程学院译、清华大学出版社、2016.3
- 机器学习，周志华，清华大学出版社
- 集体智慧编程，【美】托比 • 西格兰 (Toby Segaran)著，莫映、王开福 (译)，电子工业出版社
- 统计学习方法，李航，清华大学出版社
- 数据挖掘:概念与技术(原书第 3 版)，韩家炜、Micheline Kamber、裴健著，范明、孟小峰 (译)，机械工业出版社
- Python 机器学习:预测分析核心算法，鲍尔斯 (Michael Bowles) 著，沙嬴、李鹏 (译)，人民邮电出版社
- 统计学习基础(第 2 版)(英文)，T.黑斯蒂 (Trevor Hastie)著，世界图书出版公司
- 机器学习实战，哈林顿 (Peter Harrington)著，李锐、李鹏、曲亚东、王斌 (译)，人民邮电出版社
- Pattern Recognition and Machine Learning, Christopher M.Bishop, .Springer, 2006
- Introduction To Pattern Recognition And Machine Learning，Keinosuke Fukunaga，Academic Press; 2 edition (October 12, 1990)

CDA INSTITUTE

经管之家 CDA 数据分析研究院

2017.3